Brügelmann, Hans

Was sagen uns IQB-Bildungstrend, TIMSS, PISA und andere Ländervergleiche?

Lehren und lernen 43 (2017) 2, S. 4-9

Hans Brügelmann

Was sagen uns IQB-Bildungstrend, TIMSS, PISA und andere Ländervergleiche?

Vergleichsstudien von Fachleistungen wie TIMSS und PISA bestimmen die öffentliche Diskussion über die Qualität von Schule. Scheinbar präzise Punktwerte verführen zu Ranglisten, anhand derer die Qualität von Schule und Unterricht beurteilt wird. Übersehen werden die Unsicherheiten und Ungenauigkeiten der zugrunde liegenden Messungen. Die Bildungspolitik, vor allem aber die Schulpraxis brauchen dagegen eine Forschung, die sich der Komplexität des pädagogischen Alltags stellt.

Ende 2016 haben drei Schulleistungsstudien wieder mal für Aufregung in den Medien gesorgt.

Der IQB-Bildungstrend 2015

Als erstes kam Baden-Württemberg unter die Räder. Im „IQB-Bildungstrend 2015“ wurden vor allem die Deutsch-Leistungen von 9.-Klässler/-innen in den verschiedenen Bundesländern verglichen (Stanat u. a. 2016, im ff. nur mit Seitenzahl zitiert). Zu ihrer Erhebung wurden Tests in den Bereichen „Lesen“, „Zuhören“ und „Orthographie“ eingesetzt. Beim Vergleich der Ergebnisse wurde Baden-Württemberg unter den 16 Bundesländern auf den Rängen 10, 14 und 12 eingestuft. Nach Plätzen im oberen Bereich bei früheren Erhebungen war die Aufregung groß. Wie der Bildungsforscher *Ulrich Trautwein* im SPIEGEL (Nr. 44/2016) so malte auch *Heike Schmoll* in der FAZ vom 2.11.2016 die Schulentwicklung in Baden-Württemberg aufgrund der Daten des Ländervergleichs in düsteren Farben. Zu Recht – oder mediale Brandstiftung?

Bei den alpinen Ski-Weltmeisterschaften 2015 gewann der Schweizer Patrick Küng die Abfahrt. Er brauchte für die Strecke in Beaver Creek 1 Minute, 43 Sekunden und 18 Hundertstel. Elfter wurde sein Landsmann Didier Défago. Dank der eingesetzten Präzisionsgeräte konnte die elektronische Zeitmessung einen Rückstand von 71 Hundertstel feststellen. Nicht mal eine Sekunde langsamer – bezogen auf eine Gesamtzeit von über 100 Sekunden. Der Dritte brauchte sogar nur sieben Hundertstel mehr als der Zweite. Das sind gerade mal 0,06% mehr – für die „Ski-Kompetenz“ im Alltag belanglos. Bedeutsam ist der Unterschied nur für Wettbewerbe unter einzelnen Spitzenfahrern. Merke: Mit Präzisionsverfahren messbare Unterschiede sagen noch nichts über ihre alltagspraktische Bedeutung aus.

Kasten 1: Medaillenverteilung nach hunderstel Sekunden

Erklärungen für den angeblichen „Leistungsabfall“ hatten die Kritiker/-innen auch gleich zur Hand: Abschaffung der verbindlichen Schulempfehlung nach Klasse 4, Einführung der Gemeinschaftsschulen, Schreiben mit der Anlauttabelle im Anfangsunterricht. Und das, obwohl die Autor/-innen (S. 185) selbst ausdrücklich anmerken, „dass die Gemeinschaftsschule als neue Schulart … noch nicht berücksichtigt werden“ konnte. Und zur Grundschule sind die 9.-Klässler/-innen noch zu Zeiten der CDU/FDP-Koalition gegangen. Insofern läuft der Versuch, aus dem IQB-Trend 2015 ein Grün-Rot-Bashing zu konstruieren, leer.

Im Übrigen lohnt es, sich die Ergebnisse genauer anzuschauen. Denn was bedeuten die unterschiedlichen Rangplätze konkret?

Im Lesen erreicht Baden-Württemberg den 10. Platz mit 496 Punkten, Schleswig-Holstein auf dem 2. Platz erreicht 514 (S. 337). Sind diese 18 Punkte (oder rund 3%) Unterschied eine inhaltlich bedeutsame Differenz? Wir wissen es nicht, denn die Forscher/-innen teilen nicht mit, welchen Realunterschieden (gelesene Textmenge pro Minute, Anteil der falsch gelösten Aufgaben) diese Werte entsprechen. Das gilt auch für die Differenz zu den baden-württembergischen Ergebnissen von 2009, als das Land 521 Punkte erreichte.

Zwar übersetzen die Autor/-innen die Punktdifferenzen in zeitliche Unterschiede und sagen zum Beispiel (S. 536), in dieser Untersuchung entsprächen 20 Punkte ungefähr dem Lernfortschritt von einem Schuljahr. Aber ist der in diesem Alter wirklich bedeutsamer als die in Kasten 1 berichteten Unterschiede in der Ski-Abfahrt? Lesen durchschnittliche 9.-Klässler/-innen in derselben Zeit 2% oder 20% mehr Text als 8.-Klässler/-innen? Und lösen sie mit 15 Jahren nur 4% der Aufgaben falsch, mit 14 aber noch 10% oder auch nur 5%? Die Bedeutung der Punktunterschiede für Leistungen im Alltag bleibt also offen.

Und noch eine zweite Einschränkung: Die Ergebnisse stammen aus Stichproben, mithilfe derer die Verhältnisse in der Grundgesamtheit lediglich geschätzt werden. Eine solche Schätzung geht immer mit einem gewissen Schätzfehler einher, der es streng genommen nur zulässt einen Wertebereich anzugeben, das sogenannte Vertrauens-(„Konfidenz"-)Intervall, in dem sich der tatsächliche Wert in der Grundgesamt mit einer bestimmten Wahrscheinlichkeit befindet. In der Rechtschreibung beispielsweise liegt der tatsächliche Wert für die Grundgesamtheit aller baden-württembergischen 9.-Klässler/-innen auf Platz 12 mit 95%iger Wahrscheinlichkeit irgendwo innerhalb des Vertrauensintervalls von 491 bis 506, für die Zweitplatzierten aus Sachsen irgendwo zwischen 500 und 514 (S. 340). Die Überschneidungen der beiden Vertrauensintervalle zeigen, wie unsicher die berichteten Differenzen und damit die zugewiesenen Plätze 2 und 12 sind; denn es könnte auch sein, dass der tatsächliche Wert in der sächsischen Grundgesamtheit 501 beträgt und in der baden-württembergischen 505.

Veränderungen der Rangplätze von 2009 nach 2015 sind ähnlich wenig verlässlich zu interpretieren. Zwar erreichte die baden-württembergische Stichprobe in der Orthographie 2009 noch 516 Punkte (S. 351). Aber die Vertrauensintervalle (mit immer noch 5% Fehlerwahrscheinlichkeit) für die tatsächlichen Werte (in der Grundgesamtheit) liegen für 2015 bei 499 bis 513 und für 2009 bei 509 bis 523. Also überschneiden sich auch hier die Vertrauensintervalle. Zudem: Rangplätze sind relative Bewertungen. So sinken sie auch ohne Verschlechterung der Leistungen, wenn andere Länder zugewonnen haben.

Umso verwunderlicher das Mediengetöse. Die Autor/-innen stellen dagegen in der durchschnittlichen Rechtschreibkompetenz sachlich „keine signifikanten Unterschiede zwischen den Jahren 2009 und 2015" fest (S. 351). Warum aber soll dann im Anfangsunterricht z. B. das Schreiben mit Anlauttabellen verboten werden? Ganz zu schweigen davon, dass niemand weiß, welchen Anteil das lautorientierte Schreiben in baden-württembergischen Grundschulen tatsächlich hat (bundesweit lag z. B. Reichens Marktanteil nie über 1%). Und noch weniger ist bekannt, in welchen methodischen Kombinationen (etwa mit Grundwortschatzarbeit) es in der Regel auftritt und wie lange es durchschnittlich andauert (vgl. die Varianten in den Beiträgen zu Brinkmann 2015). Positive Effekte auf die spätere Rechtschreibung sind für das synthetisch-analytische Konstruieren von Wörtern in der Anfangsphase im deutschen wie im angelsächsischen Raum nachgewiesen (s. u. a. Richter 1992, 150ff.; National Early Literacy Panel 2008).

TIMSS 2016

Wie oberflächlich empirische Befunde in vielen Medien dargestellt und kommentiert werden, konnte man auch nach der Veröffentlichung des internationalen Grundschulvergleichs in Mathematik und den Naturwissenschaften (TIMSS) beobachten. So war auf FAZ-online am 29.11.2016 in einer dpa-Meldung zu lesen: „Deutschlands Grundschüler haben anscheinend große Probleme mit Mathematik. In diesem wichtigen Unterrichtsfach sind sie laut der Bildungsstudie TIMSS im internationalen Vergleich mit 522 Punkten (2011: 528) tief ins Mittelfeld gerutscht und liegen nun unterhalb des EU-Durchschnitts von 527 Punkten." (Hervorheb. Vf.)

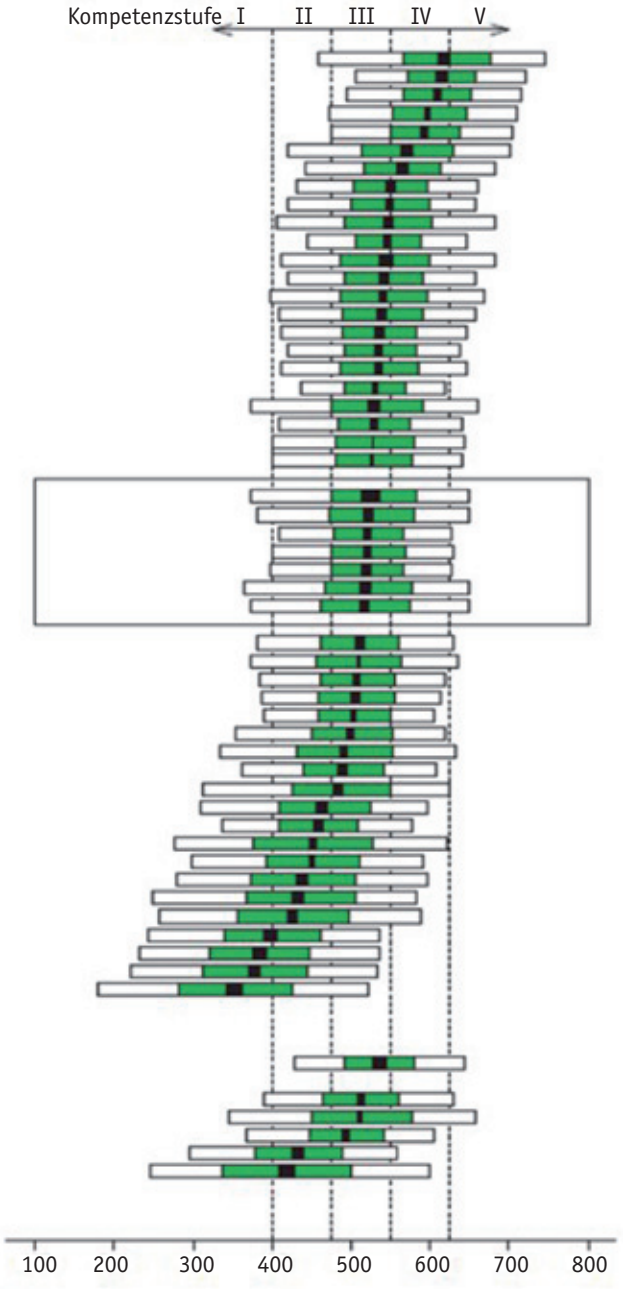
Schaut man sich die Daten genauer an, stellt man fest, dass die deutschen Viertklässler/-innen 2007 bei 525 Punkten lagen, auf diesen Basiswert bezogen also bis 2011 drei Punkte gewonnen und bis 2015 drei Punkte verloren haben. Demnach kein erkennbarer Trend, sondern Schwankungen – statistisch durchaus im Zufallsbereich für einen Trend. Zudem beziffern die Autor/-innen des TIMSS-Berichts den Lernfortschritt für ein Schuljahr auf rund 30 Testpunkte; selbst sechs Punkte Differenz (2015 vs. 2011) entsprechen also gerade mal zwei Monaten. „Große Probleme"?

Entsprechend zu relativieren ist die Differenz von 5 Punkten zum europäischen Durchschnitt, der bei 527 Punkten liegt. Denn die Verteilungen der Leistungen in den einzelnen Ländern überlappen sich erheblich. Dass Deutschland nicht mehr über dem EU-Durchschnitt liegt, kann zudem nicht einfach als Folge schlechterer Leistungen in unserem Land interpretiert werden („tief … gerutscht"), sondern ist zumindest auch einer Verbesserung der Testergebnisse in anderen Ländern geschuldet: der europäische Durchschnitt lag 2011 bei 519 Punkten.

Vor allem aber wurde in den Medien kaum zur Kenntnis genommen, dass sich die deutsche Schülerpopulation von 2011 bis 2015 bedeutsam verändert hat, so dass Veränderungen nicht ohne weiteres auf den Unterricht zurückgeführt werden können. In den Worten der Autor/-innen des TIMSS-Berichts: „In Mathematik sind unter Berücksichtigung von Veränderungen in der Schülerschaft die durchschnittlichen Leistungen von dem Jahr 2007 zu dem Jahr 2011 statistisch signifikant um 11 Leistungspunkte gesunken und von TIMSS 2011 zu 2015 statistisch signifikant um 8 Punkte gestiegen. Damit wurde in TIMSS 2015 wieder das Leistungsniveau von TIMSS 2007 erreicht." (S. 375, Hervorheb. Vf.)

Das Bild ist also viel komplizierter als in der öffentlichen Diskussion verhandelt. Und: Vorschnelle Ursachen- und Schuldzuweisungen können zu fatalen Fehlschlüssen bei der Entscheidung für die erforderlichen Maßnahmen führen – wie sich schon oben bei den Fehlinterpretationen der IQB-Trendstudie gezeigt hat.

Testleistung der Schülerinnen und Schüler im internationalen Vergleich – Gesamtskala



Teilnehmer	M	(SE)
Singapur	618	(3.8)
Hongkong	615	(2.9)
Republik Korea (Südkorea)	608	(2.2)
Taiwan	597	(1.9)
Japan	593	(2.0)
Nordirland	570	(2.9)
Russische Föderation	564	(3.4)
Norwegen (5. Jgst.)	549	(2.5)
Irland	547	(2.1)
England	546	(2.8)
Belgien (Fläm. Gem.)	546	(2.1)
Kasachstan	544	(4.5)
Portugal	541	(2.2)
USA	539	(2.3)
Dänemark	539	(2.7)
Litauen	535	(2.5)
Finnland	535	(2.0)
Polen	535	(2.1)
Niederlande	530	(1.7)
Ungarn	529	(3.2)
Tschechische Republik	528	(2.2)
VG OECD	528	(0.5)
VG EU	527	(0.6)
Bulgarien	524	(5.3)
Zypern	523	(2.7)
Deutschland	522	(2.0)
Slowenien	520	(1.9)
Schweden	519	(2.8)
Serbien	518	(3.5)
Australien	517	(3.1)
Kanada	511	(2.3)
Internationaler Mittelwert	**509**	**(0.4)**
Italien	507	(2.6)
Spanien	505	(2.5)
Kroatien	502	(1.8)
Slowakei	498	(2.5)
Neuseeland	491	(2.3)
Frankreich	488	(2.9)
Türkei	483	(3.1)
Georgien	463	(3.6)
Chile	459	(2.4)
Vereinigte Arabische Emirate (VAE)	452	(2.4)
Bahrain	451	(1.6)
Katar	439	(3.4)
Iran	431	(3.2)
Oman	425	(2.5)
Indonesien	397	(3.7)
Saudi-Arabien	383	(4.1)
Marokko	377	(3.4)
Kuwait	353	(4.6)
Benchmark-Teilnehmer		
Québec, Kanada	536	(4.0)
Ontario, Kanada	512	(2.3)
Dubai, VAE	511	(1.4)
Norwegen (4. Jgst.)	493	(2.3)
Buenos Aires, Argentinien	432	(2.9)
Abu Dhabi, VAE	419	(4.7)

Kasten 2: Mittelwertsunterschiede und Vertrauensintervalle, Überlappende Leistungsverteilungen bei TIMSS 2015 (aus Abb. 3.5 in Wendt u. a. 2016, 107)

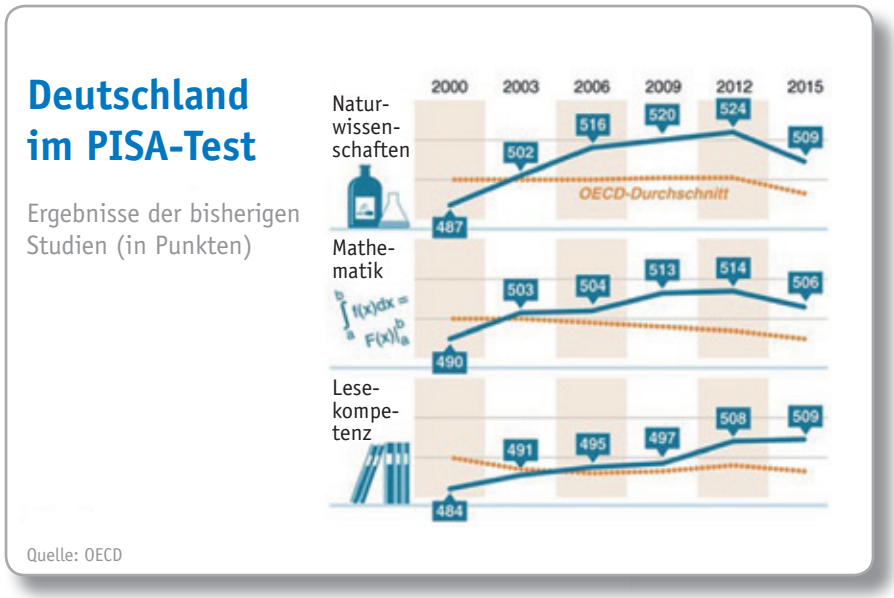
PISA 2015

Und dann kam PISA – zum sechsten Mal seit 2000. Nach den Erfolgsmeldungen der letzten Runden, in denen die deutschen 15-Jährigen jeweils Punkte gewonnen und hier und da Plätze gut gemacht hatten, dieses Mal „Stabilisierung auf hohem Niveau“ (so KMK-Präsidentin Claudia Bogedan, Bremen): Im Lesen +1 Punkt, in Mathematik –8 und in den Naturwissenschaften –15 Punkte. Die rücksichtsvolle Bewertung erstaunt, wurden doch in den Vorjahren schon Zuwächse von 4 oder 9 Punkten zu „Fortschritten“ hochgejubelt und als „Beweis“ für eine erfolgreiche Bildungspolitik verkauft: die Einführung von Bildungsstandards und regelmäßigen Tests habe die „Qualität“ der Schulen gesteigert. Nun heißt es, solche Differenzen seien „statistisch nicht signifikant“.

Das stimmt, aber dann muss man sich das Ganze doch etwas grundsätzlicher anschauen. Bekanntlich werden auch kleine Unterschiede „statistisch signifikant“, wenn

die Stichproben nur groß genug sind. Bei PISA umfassen sie 5.000 bis 10.000 Schüler/-innen pro Land. Wenn dann Unterschiede von 5 bis 10 Punkten (je nach Messfehler) oft „nicht signifikant“ werden, können sie auch inhaltlich nicht besonders bedeutsam sein. Bei PISA-2000 konnte schon die Lösung einer einzigen Aufgabe zusätzlich zu 20 Punkten Zugewinn führen. Die Umrechnung der Punkte auf eine 500er-Skala lässt selbst kleine Unterschiede groß erscheinen – visuell zusätzlich verzerrt, wenn in den Kurven die unteren 480 Punkte „abgeschnitten“ werden (s. Kasten 3).

Und in der Tat erreichen selbst 10–20 Punkte auf der PISA-Skala in der Regel nicht die Schwelle auch nur „kleiner Effektstärken“, die für Mittelwertsunterschiede bei d = 0.2 bis 0.3 angesetzt werden (s. Kasten 4).



Kasten 3: Unterschiede ohne Basisbezug (dpa/ FAZ-online v. 6.12.2016)

Rangplätze in einer Stichprobe sagen nicht viel über reale Abstände aus. Darum muss man die Unterschiede zwischen den Punktwerten in mehrfacher Hinsicht prüfen.

1. Ist angegeben, ob Punktdifferenzen zwischen Gruppen „statistisch signifikant“ sind? Nur dann sind die zwischen den Stichproben gefundenen Unterschiede mit großer Wahrscheinlichkeit auch zwischen den entsprechenden Grundgesamtheiten (bspw. alle 15-jährigen Schüler/-innen Deutschlands und Finnlands) zu finden. [Technische Anm.: Die realen Punktwerte liegen mit (immer noch 5%igem Fehlerrisiko) im Bereich von rund +/– zwei Standardfehlern um die Werte in den Stichproben (in dem „Vertrauensintervall“).] Die deutschen Schüler/-innen könnten demnach bei einer Wiederholung von PISA statt auf Platz 10 (bezogen auf die OECD-Länder) wegen der Messfehler auch auf Platz 6 oder Platz 13 landen, in 5% der Fälle sogar außerhalb dieses Bereichs (und das auch, falls sich die Werte anderer Länder ebenfalls verändern). Schon die Plätze auf den Ranglisten sind also sehr unsicher …
2. Sind Unterschiede statistisch signifikant, sind sie damit aber noch nicht inhaltlich bedeutsam. Denn bei gleichen Durchschnittswerten spielt es eine Rolle, wie breit die Einzelwerte innerhalb einer Gruppe um den Mittelwert streuen. Scharen sich die Werte in beiden Gruppen eng um den Durchschnitt, ist der Unterschied zwischen den Mittelwerten von zwei Gruppen trennschärfer, als wenn sich beide Verteilungen breit überlappen. Technische Anm.: Ein Maß für diese Trennschärfe ist die Effektstärke – meist berechnet als Differenz der beiden Mittelwerte geteilt durch die gemittelte Streuung in den beiden Gruppen (zur einfachen Berechnung aus den üblichen Tabellenwerten s. das Formular unter www.soerenwallrodt.de). Dieser Wert sollte mindestens 0.2–0.3 (Schwelle für einen sogar nur „schwachen Effekt“), besser sogar mehr als 0.5 („mittlerer Effekt“) erreichen. Differenzen selbst von 10 Plätzen in den hier berichteten Studien erreichen meist nur Werte von 0.1 bis 0.2.
3. Letztlich kommt es aber auf das fachlich begründete Urteil an, welche Bedeutung man Punktdifferenzen in persönlicher Verantwortung zuerkennt. Werden Originaldaten angegeben, kann man die Bedeutung von Unterschieden für den Alltag selbst abschätzen: Macht z. B. die eine Gruppe auf 1.000 Wörter 81 Rechtschreibfehler und die andere 79, spielt die Differenz von 2 Fehlern für den Alltag sicher keine große Rolle. Oft werden aber – wie bei PISA oder TIMSS – statt der Originaldaten nur umgerechnete Werte berichtet. Manchmal veranschaulichen die Autor/-innen die Bedeutung von Unterschieden durch den Vergleich mit Lernfortschritten über die Zeit hinweg. Aber auch dann muss man aufpassen: Im Lesen verbessert sich die Leistung über ein Schuljahr hinweg auf Klassenstufe 1/2 erheblich, auf Klassenstufe 9/10 dagegen nur noch wenig. Geht es andererseits um neue Inhalte (z. B. in Mathematik), kann auch auf Klassenstufe 9/10 ein Schuljahr einen bedeutsamen Unterschied darstellen.

Mehr zu den technischen Details bei Lind (2016): https://www.uni-konstanz.de/ag-moral/pdf/Lind-2016_Effektstaerke-Vortrag.pdf

Kasten 4: Wann sind Unterschiede in Testergebnissen „bedeutsam“ für Bildungspolitik und für den pädagogischen Alltag?

Nachdenklich stimmt auch, dass Korea von 2012 bis 2015 in den Naturwissenschaften 30 Punkte verloren hat, die Schüler/-innen also binnen drei(!) Jahren um ein ganzes Schuljahr schwächer geworden sein sollen. Schweden andererseits hat von 2003 bis 2012 insgesamt 31 Punkte verloren, dann aber bis 2015 schon wieder 16 Punkte aufgeholt, was angeblich einem halben Schuljahr entspricht. Werden in solchen Schwankungen tatsächlich reale Veränderungen abgebildet?

Es gibt auch noch andere Erklärungen. So mussten die Aufgaben bis 2012 mit Bleistift auf Papier bearbeitet werden – 2015 aber am Computer. In einem Vergleich beider Formen stellten *Robitzsch* u. a. (2016) fest, dass deutsche Schüler/-innen bei denselben Aufgaben am PC deutlich (um 10 und mehr Punkte) schlechtere Ergebnisse erzielten als auf dem Papier. Dafür, dass dieser Medienwechsel einen gewichtigen Einfluss auf die Leistungsvergleiche hat, spricht auch ein Ergebnis aus Österreich: Dort haben sich im Lesen die Mädchen verschlechtert, die Jungen etwas verbessert (Nimmervoll 2016).

Wenn aber schon der Wechsel der Aufgabenform solche Unterschiede ausmachen kann – was bedeutet das erst für die Aussagekraft der künstlichen Testsituation für Leistungen unter Alltagsbedingungen? Und welche Bedeutung hat der Zeitdruck für den Abruf vorhandener Kompetenzen, welche Rolle spielt dabei z. B. die unterschiedliche Textlänge der Aufgabe in verschiedenen Sprachen, wie stark überlagern zudem die Leseanforderungen der textgebundenen Aufgabenstellung die mathematischen und naturwissenschaftlichen Leistungen? Viele Fragezeichen.

So ist auch denkbar ist, dass die Leistungszuwächse vorher, also nach 2000 auf eine wachsende Vertrautheit der Schüler/-innen mit den Aufgabenformaten und deren stärkere Nutzung im Unterricht zurückzuführen sind.

Außerdem verändern sich die Maßstäbe für die Bewertung erreichter Punkte, nämlich die Mittelwerte von EU, OECD und internationaler Stichprobe, von Termin zu Termin, weil nicht immer dieselben Länder teilnehmen, so dass sich Rangplätze selbst bei gleich bleibenden Ergebnissen verändern können.

Das PISA-Ranking und das mit ihm verbundene Verständnis von Forschung ist demnach hoch problematisch (ausführlicher: Jahnke/Meyerhöfer 2006, Brügelmann 2015, Dammer 2015).

Das Desiderat: alternative Lehr-Lern-Forschung

Interessanter als die Ergebnisse solcher Studien sind Fragen, die sie aufwerfen: So erreichen bei PISA-2015 die deutschen Schüler/-innen in den Naturwissenschaften – bezogen auf Unterrichts- und Hausaufgabenzeit – mehr Punkte als alle anderen Länder außer Finnland. Ein Indiz für besonders effektives Lernen, für besonders erfolgreiche Lehrer/-innen? Und anders als erwartet sind sie im Fachwissen besonders gut – im methodischen Denken eher schwach. Letzteres beherrschen angeblich die asiatischen Schüler/-innen besser, deren Schulen immer wieder reiner Drill vorgeworfen wird.

An der Untersuchung solcher Überraschungen sollte Bildungsforschung ansetzen. Und vor allem sollte sie untersuchen, wie es manchen Schulen gelingt, selbst unter schwierigen Bedingungen erfolgreich zu arbeiten. Aber das erfordert einen anderen Stil von Forschung: Lernbiographien von Schüler/-innen, Beobachtungen der Interaktionen im Unterricht aus verschiedenen Perspektiven, Dokumentation und Analyse von Umsetzungsvarianten desselben Programms/derselben Methode, Fallstudien der Entwicklung von Schulen und Lehrer/-innen. Die nächste PISA-Olympiade kann dafür noch zehn Jahre warten.

Dieser Artikel erscheint mit freundlicher Genehmigung der Redaktion als Nachdruck aus: Grundschule aktuell, Nr. 137, Februar 2017.

Literatur

Brinkmann, E. (Hrsg.): Rechtschreiben in der Diskussion – Schriftspracherwerb und Rechtschreibunterricht. (Beiträge zur Reform der Grundschule, Bd. 140) Frankfurt/M.: Grundschulverband 2015.

Brügelmann, H.: Vermessene Schulen – standardisierte Schüler. Zu Risiken und Nebenwirkungen von PISA, Hattie, Vera & Co. Weinheim/Basel 2015.

Dammer, K.-H.: Vermessene Bildungsforschung. Wissenschaftsgeschichtliche Hintergründe zu einem neoliberalen Herrschaftsinstrument. Baltmannsweiler 2015.

Jahnke, T./Meyerhöfer, W. (Hrsg.): Pisa & Co. Kritik eines Programms. Hildesheim 2006.

National Early Literacy Panel (Ed.): Developing early literacy: A scientific synthesis of early literacy development and implications for intervention. (National Institute for Literacy & The Partnership for Reading) Jessup, Maryland 2008.

Nimmervoll, L.: Statistiker hinterfragt Österreichs PISA-Verschlechterung. In: Der Standard v. 8.12.2016. Download: http://derstandard.at/2000048953546/Statistiker-hinterfragt-Oesterreichs-Pisa-Verschlechterung

Reiss, K., u. a. (Hrsg.): PISA 2015. Eine Studie zwischen Kontinuität und Innovation. Münster/ New York 2016.

Richter, S.: Die Rechtschreibentwicklung im Anfangsunterricht und Möglichkeiten der Vorhersage ihrer Störungen. (Phil. Diss. Universität Bremen) Hamburg 1992.

Robitzsch, A., u. a.: Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. Eine Skalierung der deutschen PISA-Daten. 2016. Online am 6.12.2016: http://econtent.hogrefe.com/doi/full/10.1026/0012-1924/a000177

Stanat, P., u. a. (Hrsg.): IQB-Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich. Münster/New York 2016.

Wendt, H., u. a. (Hrsg.): Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. Münster /New York 2016.

Prof. em. Dr. Hans Brügelmann
Fachreferent für Qualitätsentwicklung
im Grundschulverband
hans.bruegelmann@gmx.de

Albrecht Wacker, Reinhold Funke, Rolf Göppel, Karl-Heinz Dammer, Hans-Werner Huneke, Karin Vogt, Andreas Müller-Hartmann, Anne Sliwka

Zur öffentlichen Diskussion des IQB-Bildungstrends 2015

Der im Oktober 2016 veröffentlichte IQB-Bildungstrend für sprachliche Kompetenzen im Vergleich der Bundesländer bescheinigte Baden-Württemberg unterdurchschnittliche Leistungen. Unmittelbar danach folgte in Zeitungen und Magazinen eine rege Diskussion zu möglichen Ursachen. Eine interdisziplinäre Arbeitsgruppe der Pädagogischen Hochschule Heidelberg, bestehend aus Fachdidaktikern der Fächer Deutsch und Englisch sowie Bildungswissenschaftlern, unterzog die vorgebrachten Argumente einer genaueren Prüfung. In ihrer hier nachfolgenden Stellungnahme geht es auch generell um den öffentlichen Umgang mit Ergebnissen der quantitativen Bildungsforschung.

Am 28. Oktober 2016 wurde der IQB-Bildungstrend veröffentlicht, der zum zweiten Mal sprachliche Teilkompetenzen von Schüler/-innen am Ende der 9. Klassenstufe über alle Bundesländer hinweg ausweist. Getestet wurden darin ausgewählte Kompetenzen aus den Fächern Deutsch (Lesen, Zuhören, Orthografie), sowie Englisch und Französisch (jeweils Lese- und Hörverstehen). In der Studie werden die Leistungen der Schüler/-innen sowohl in kriterialer Perspektive (in Bezug auf das Erreichen von Bildungsstandards) als auch in sozialer Perspektive (in Bezug auf einen Vergleich der Bundesländer) und in zeitlicher Perspektive (in Bezug auf einen Vergleich mit der Studie von 2010) ausgewiesen.

Für das Bundesland Baden-Württemberg ergaben sich vor allem im Fach Deutsch ungünstige Ergebnisse, weil deutlich wurde, dass in den Teilkompetenzen des muttersprachlichen Lesens und Zuhörens der Anteil der Jugendlichen, die den Regelstandard erreichen oder übertreffen, zwischen den Jahren 2009 und 2015 signifikant zurückgegangen und umgekehrt der Anteil der Schülerinnen und Schüler, die den Mindeststandard verfehlen, signifikant gestiegen ist.

Unmittelbar nach Veröffentlichung der Studie (Oktober 2016) führten Journalisten und Bildungsforscher zahlreiche potenzielle Ursachen für das schlechte Abschneiden an, die zumeist die Systemebene und Fragen der Schulstruktur und Bildungsorganisation, sowie Fragen der Professionalisierung von Lehrpersonen in den Blick nahmen.

Die nachfolgende Stellungnahme der Heidelberger Arbeitsgruppe betrifft die Deutung der Befunde generell (1.), das Problem des öffentlichen Umgangs mit Forschungsergebnissen (2.) sowie in der öffentlichen Debatte vorgebrachten „Erklärungs-“Ansätze (3.).

Zur Art und der Reichweite der Befunde

Aus Sicht der Arbeitsgruppe ist eine klare Benennung von Ursachen oder gar von Schuldzuschreibungen, wie sie bisher öffentlich stattfand, auf der Grundlage der Daten nicht möglich, weil die Studie allein von ihrer Anlage her kaum Rückschlüsse auf Ursachen erlaubt, die für den Rückgang der Schülerleistungen in Baden-Württemberg verantwortlich sein könnten. Auch lassen die getesteten Kompetenzfacetten keinen Schluss auf die Beurteilung der Leistungen in einem Fach als Ganzem zu, da beispielsweise die in standardisierten Tests gemessenen Kompetenzen in sprachlichen Fächern deren konstitutive kommunikative Aspekte (oder auch in Fremdsprachen die bedeutsamen interkulturellen Aspekte) nicht abzubilden vermögen.

Deshalb erscheinen für die Arbeitsgruppe viele Argumente, die in den ersten Pressemeldungen und -kommentaren angeführt wurden, recht beliebige Einschätzungen zu sein, von denen manche zwar eine gewisse Plausibilität haben mögen, aber aus den Studien heraus nicht wissenschaftlich schlüssig abgeleitet werden können. Schon gar nicht taugen sie zur Begründung eines Krisenszenarios.